Title
The Realism of Precipitation Extremes in High-Resolution Gridded Datasets: A Case Study over California

Author
Grieco, Matthew

Publication Date
2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

The Realism of Precipitation Extremes in High-Resolution Gridded Datasets: A Case Study over California

A thesis submitted in partial satisfaction of the

requirements for the degree Master of Science

in Atmospheric and Oceanic Sciences

by

Matthew Brian Grieco

2018

ABSTRACT OF THE THESIS

The Realism of Precipitation Extremes in High-Resolution Gridded Datasets: A Case Study over California

by

Matthew Brian Grieco

Master of Science in Atmospheric and Oceanic Sciences

University of California, Los Angeles, 2017

Professor Alexander Hall, Chair

There is a growing need for high-resolution, spatially complete meteorological data. These data are utilized within weather, climate, ecological, and environmental research. With so many different gridded datasets available, it is unclear which is best for a given application. There are few comprehensive studies that have examined the strengths and weaknesses of gridded datasets, and they have mainly identified flaws or differences without understanding how these differences arise from the methodologies or suggesting ways to improve. Precipitation extremes are especially difficult to capture in gridded data. Here we assess precipitation extremes of five high-resolution gridded datasets over California to by comparing with daily station data and interpret the results in the context of each dataset's methodology. Multiple statistics of extreme precipitation were considered, reflecting intensity, frequency, and duration. Large differences are found both between the gridded datasets and relative to the station data. Maximum single day precipitation is underestimated in nearly all datasets, and precipitation frequency is severely overestimated in many cases. Datasets differ most notably in magnitude and precipitation occurrence in mountainous and coastal regions. The errors of these gridded datasets can vary significantly when compared to station data; one dataset within this study gives a 54% error for consecutive precipitation frequency. The results of this assessment are likely to be useful for users of gridded datasets looking to select a dataset appropriate for their research. They could also aid gridded dataset creators in improving existing products or building new ones.

The thesis of Matthew Brian Grieco is approved.

J. David Neelin

Rong Fu

Yongkang Xue

Alexander Hall, Committee Chair

University of California, Los Angeles

2018

Table of Contents

List of Figures and Tables

Acknowledgements

I would like to thank my incredible advisor, Professor Alex Hall, for supporting me through the entirety of this project. I am very grateful for his patience and encouragement, not only on the research and results of the project, but within my own academic and career aspirations as well. While graduate school and research can be gritting tasks, I always knew that if any issues arose, I could immediately go to him for guidance, advice and support; and for that I am immensely thankful.

I would also like to give a very special thank you to Daniel Walton, who I, rightly so, call my "pseudo-advisor" for this project. From the beginning, he supported my work and promoted the rigorous investigations necessary for the success of the project. Being able to bounce ideas off each other and work together in real time –even through Skype, truly made the researching and writing experiences much more enjoyable. I would never have been able to complete this research nor thesis without his expertise and assistance along the way.

Finally, coding and research, no matter the level of familiarity, can be highly intimidating. However, being able to work someone who knows the specifics makes the seemingly insurmountable tasks much more manageable. Therefore, I'd also like to thank Baird Langenbrunner for his generosity and aid with the project's details. For any trouble with graphs, figures or coding specifics, he was always ready to help any way he could, and for that I express my sincerest gratitude.

1. Introduction

1.1 Gridded Datasets Definition and Background

Accurate historical meteorological data are key to many hydrologic, weather system, and climate models. While quality-controlled weather stations are generally understood as our best primary source precipitation data, they are point measurements, and spatially complete data is necessary for many applications (Behnke et al. 2016a; Abatzoglou 2013). Many gridded datasets –defined here as spatially complete data on regular grid– exist but few studies have attempted to evaluate their systematic errors (Behnke et al. 2016a). Gridded datasets vary in their gridding algorithms, parent or training station data, temporal adjustments, and resolution. Having high resolution (<12 km) is useful for better capturing the realistic distribution of parameters that vary with topography; such as precipitation. Gridded datasets have been utilized within fields of hydrology and agriculture (Thornton et al. 1997, Mote et al. 2005, Abatzoglou 2013), as well as to evaluate regional climate models (Caldwell et al. 2009, Walton et al. 2015) and even to train statistical models (Pierce et al. 2014). It is widely acknowledged that gridded datasets have flaws and their meteorological products are intrinsically uncertain due to their approximations and methodological choices (Daly et al. 2008, Newman et al. 2015, Behnke et al. 2016a, Walton et al. 2018). Some factors which lead to this uncertainty are interpolation methods, sparse or questionable data, and unrealistic reliance on physical properties such as topography (Newman et al. 2015). Walton et al. (2018) performed a comprehensive comparison of temperature in gridded datasets, but differences in precipitation gridded datasets may be even more severe. Behnke et al. (2016a) supports this assumption as their findings show differences in precipitation extremes to be especially large.

1.2 Known Issues from the Literature

Although there have been multiple studies revealing that the collection of gridded datasets are unique amongst each other and have important differences compared to the most trusted station data (Ensor and Robeson 2008, Mannshardt-Shamseldin et al. 2010, Mizukami and Smith 2012, Gervais et al. 2014, Prein and Gobiet 2016, Behnke et al. 2016a, Henn et al. 2018) much reliance and trust is put into these gridded datasets. Most users are interested temperature and precipitation variables, especially for areas where station density is low (Bajamgnigni et al. 2016). Even for places with high station density, well-known errors exist. One such error may be from the raw data itself, which can have systematic errors. For precipitation gauges, these errors include precipitation undercatch due to the turbulent wind field created around the gauge's catch opening, precipitation underestimation from water's adhesive properties to the gauge sides, and evaporation of the water before the exact measurement is taken (Adam and Lettenmaier 2003; Legates 1987; Sevruk 1982). The underestimation from the gauges at many stations is not unique to liquid precipitation. Pan et al. (2003) found that snowfall is also underestimated because of the same issues described above. A few of studies have supported these findings, stating that within mountainous areas of Europe, station gauges can generally be corrected 3 to 20% for liquid precipitation, and between 40 and 80% (depending on gauge shielding) for snow, because of these well-known physical issues (Førland and Institutt, 1996; Goodison et al., 1997).

As stated above, while it is understood that gridded datasets house many imperfections, precipitation errors from these gridded datasets, to the best of our knowledge, have not been analyzed with as much scrutiny as their gauge-based counterparts. Nonetheless, some errors are prevalent within the literature. Through an analysis of gridded precipitation datasets within the central U.S., Ensor and Robeson (2008) revealed a significant increase in days with low precipitation and conversely, a large reduction of frequency of days with extreme precipitation, when comparing against station data. They also noted, that even with this variation amongst precipitation days, the time series between the station data and grid points are well correlated. Eum et al. (2014) also found various quantitative differences among precipitation datasets from their study domain in Canada. They found quantitative inconsistencies (minor and major) amongst gridded datasets, particularly in within mountainous regions, and concluded that these inconsistencies lead to an intensified differencing among output for hydrologic modeling. Tozer et al. (2012) also concluded that clear differences in hydrologic run-off modeling exist amongst gridded precipitation datasets, particularly in very dry or very wet areas. These inconsistencies lead to vastly different runoff values and feedbacks in hydrologic models. In general, Behnke et al. (2016a) highlighted a rather consistent underestimation of precipitation extremes, with some areas exhibiting biases of more than 50% for high precipitation events in all the gridded datasets within their study. Others, like Hofstra et al. (2010), have hypothesized this to be a common error due to the interpolation methods utilized to create these gridded datasets. Hofstra et al. (2010) noted that interpolation methods generally smooth spatial variability as a correction factor, leading to non-realistic results, and in the case of precipitation, underestimated extremes. Modeling exercises have shown that the differences in precipitation extrema between station and gridded data is extremely large; some stations give return values for these extremes two or three times that of the gridded data (Mannshardt-Shamseldin et al. 2010). As one would expect, estimating precipitation in complex topography is challenging (Gutmann et al. 2012, Henn et al. 2018). Henn et al. (2018) found that within the Western U.S., the greatest quantitative differences among datasets exist in high elevation regions, specifically coastal mountain ranges, like the Sierra Nevada Mountains. Their analysis found precipitation differences of 200 mm/year or greater, or 5-60%, within the complex terrain of this area, as well as extremely variable multi-year trends with the gridded datasets (Henn et al. 2018).

Interpolation methods are a highly important factor in determining the differences between gridded datasets and station data. There are many different methods to generating a spatially complete precipitation field from point measurements. Interpolation methods may rely on: station to grid point distance, station density, or topographic features like elevation, slope, and ground type. Station distance is typically a key factor in the interpolation methods. Methods like simple inverse distance weighting (Cressman 1959; Shepard 1984), schemes that see optimize mean-square errors (Gandin 1965), kriging methods (identifying spatial data as a function of distance and direction) (Journel and Huijbregts 1978), and even blending techniques together (Johns et al. 2003) all heavily rely on station to grid point distance. All methods employed to create gridded datasets (including the ones named above) do come with intrinsic errors for any variable chosen. Thus, comparison amongst datasets with different interpolation methods can help to solidify known flaws and reveal others. In terms of precipitation, station density is also an extremely important factor for the interpolation methods. As station density decreases, the interpolation methods must use stations farther and farther away from current grid point, possibly resulting in a grid point's precipitation calculated from stations with very dissimilar precipitation distributions (Gervais et al. 2014). A study by Daly (2006) suggested that objectively, inland grid points (>100 km from coast) are much easier to represent as opposed to grid points influenced by coastal and topographical features. Daly also suggests that grid points with more coastal or topographical influences require higher station density to achieve equal accuracy.

1.3 Objectives of Study

The above studies show that some assessment of gridded datasets have been made. However, these studies have been (1) too spatially vast, generalizing flaws to a whole country with immensely different meteorological and/or topographic influences, (2) have only a small number of datasets with similar interpolation methods, or (3) do not necessarily assess the reasoning behind the flaws that arise from careful analysis (i.e., because of algorithms, topography, fit to station data, etc.). Also, for gridded dataset users, choosing a single dataset from the many that exist without knowledge of differences and flaws they possess can be overwhelming. Therefore, the one goal of this paper is to comprehensively evaluate the suitability of available gridded datasets for studying extreme precipitation. The author hopes this will aid gridded dataset users in the selection process, but also spur gridded dataset creators to improve existing products and/or build new ones. This paper is structured as follows. In Section 2, all gridded datasets and the station data initially analyzed in this study are explained in detail. This section highlights each gridded dataset's important characteristics, such as interpolation methods, resolution, parent station data utilized, and (if any) spatial and/or temporal corrections made (see Table 1 and data section for succinct gridded dataset information). Section 3 describes, in detail, the methods employed to assess the differences in precipitation extremes between these datasets. Section 4 houses the results from this analysis, described in high quantitative detail. The final section (Section 5) gives potential reasons behind the precipitation differences identified within the paper, as well as discussions and recommendations about these findings as a whole

2. Data

2.1 GHCND (station data)

Station data are our best primary source for precipitation data and are used here to identify biases and errors in the gridded datasets. Here we use a quality-controlled subset of the Global Historical Climatology Network – Daily station data (GHCND; Menne et al. 2012), provided by Behnke et al. (2016b) via the Dryad data package (http://dx.doi.org/10.5061/dryad.7tv80). This dataset contains 3855 stations over the contiguous U.S. with the temperature and precipitation values for at least 83% of the 1981-2010 period, making it appropriate for our study. Because this study addresses extreme precipitation, it is important that days having the most extreme precipitation are not missing from the station record. Therefore, we added further restrictions; stations must have daily precipitation data for at least 90% of the 1981-2010 period, and of the stations with data that passed this criterion, less than 10% of that remaining data must have non-trivial values in between this same temporal period. Overall, 221 California stations passed this threshold criteria and were utilized in the analysis (Fig 1). While a handful of stations in the central Sierra Nevada mountains, near the California-Nevada border, passed the imposed criteria, individual case analysis showed that the data from these stations were incorrectly recorded, and therefore not used within the analysis. In total, 13 stations from the GHCND dataset were omitted from further extreme and statistical analysis. The number of stations mentioned above –221 California stations– already accounted for these omitted 13 stations.

2.2 WRF

The first dataset analyzed in this study was produced by dynamically downscaling 32-km resolution NCEP North American Regional Reanalysis (NARR; Mesinger et al. 2006) for the period 1981-2010 using the Weather Research and Forecasting model v3 (WRF; Skamarock et al. 2008) performed by Walton et al. (2018). Under this setup, WRF is forced at the lateral and ocean surface boundaries by NARR. WRF is setup with three nested grids of 27-km, 9-km, and 3-km resolution. This study focuses on the middle domain (D2) with 9-km resolution that covers the entire state of California (Fig 1). For a full description of the model setup, see Walton et al. (2018). WRF is the only dynamically downscaled dataset in this analysis, while the others are station-based. Thus, an important aspect of this work will be characterizing the differences between WRF and the other datasets.

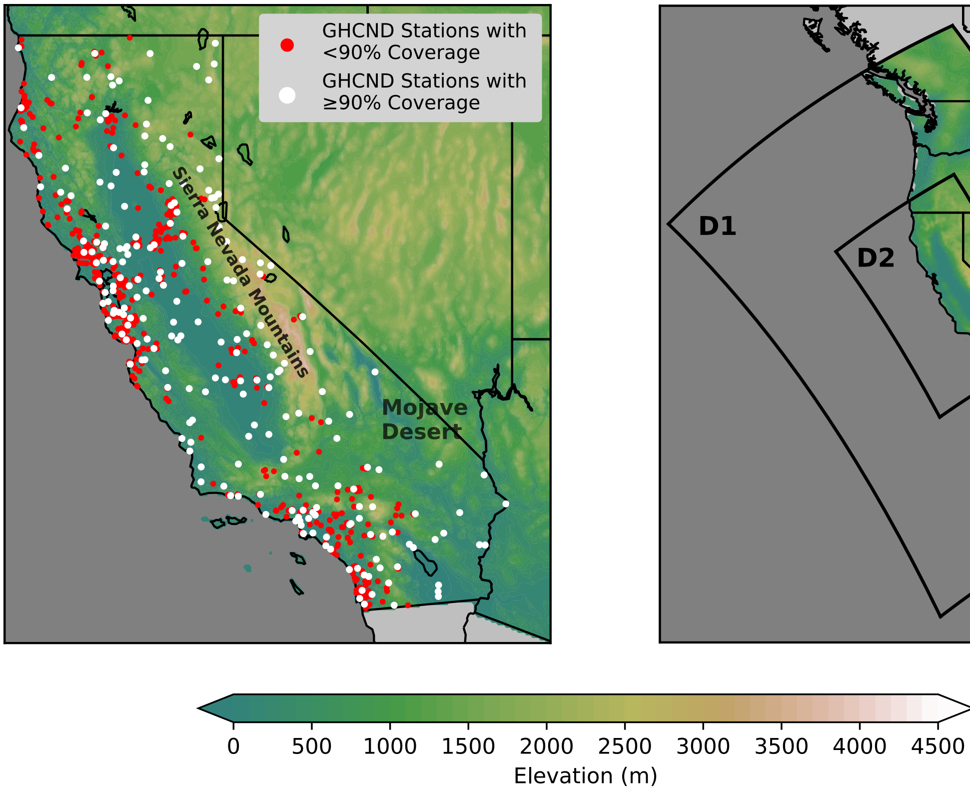

Figure 1: California Stations, Topography and Domains. (LEFT) Topography for the Western U.S., GHCND stations within California with <90% time coverage (red), and GHCND stations within California with ≥90% time coverage (white); i.e., stations utilized in this study. (RIGHT) Western U.S. topography within WRF domain "D1", and WRF domain "D2". D2 is the downscaled domain utilized within this study.

2.3 Livneh

The third dataset (Livneh et al. 2015; henceforth referred to as Livneh) is a station-based gridded dataset of daily minimum temperature, maximum temperature, and precipitation. The data were downloaded from https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.nodc:0129374. This 1/16° (approximately 6-km) resolution gridded dataset covers the contiguous U.S., Mexico, and portions of southern Canada over the period between 1950 and 2013. This dataset is an updated version of the Livneh et al. 2013 dataset, with improvements upon spatial extent, orographic precipitation representation over Canada and Mexico, and transboundary discontinuities. The precipitation values are adjusted so that the 1981-2010 precipitation monthly climatology matches PRISM (see part d of this section). The gridding is performed via the synergraphic mapping system (SYMAP, Shepard 1984), where the grid point in question is given a value using a weighted average of four nearest stations. Weighting of the nearby stations is determined by a combination of inverse distance weighting (IDW) and down-weighting those that are in a similar direction as other nearby stations. For details of the implementation, see Maurer et al. (2002) and Livneh et al. (2013).

2.4 PRISM

The Parameter-elevation Relationships on Independent Slopes Model (PRISM; Daly et al. 1994, Daly et al. 2008) is a used to develop spatial climate datasets for precipitation, maximum and minimum temperature and several other variables over the contiguous U.S. There are many different versions available, with different temporal frequency, resolution, and extent of period covered; several versions of lower frequency data (monthly or annual) have data from 1895 to the present. There are two resolutions from http://prism.oregonstate.edu (the PRISM Climate Group at Oregon State University) with daily temperature and precipitation variables, the 800 m and ~4-km (2.5 min). The free and accessible 4-km resolution AN81d dataset from the period 1981 to the present was utilized in this study (version: 14.1-20140502-1000; downloaded on February 21, 2017). PRISM uses an elevation regression function calibrated at each grid point using a moving window approach. Station weights incorporate physically-based factors such as: distance, elevation, vertical layer, topographical facet, station density, position, topographic features, coastal proximity and effective terrain height. PRISM utilizes data from many networks, including RAWS, COOP, CDEC, Agrimet, SNOTEL, NCRS, EC (Canada), and more. It is important to note that the station data are scrutinized through multiple quality control measures, but the final data are not calibrated to achieve any temporal homogeneity. For more information about the gridding algorithm and the station networks used, visit http://prism.oregonstate.edu.

2.5 Daymet

Daymet (Thornton et al. 1997; Thornton et al. 2017) provides daily hydrological variables for North America (including Hawaii, Puerto Rico and Bermuda) on a 1-km resolution grid for the period 1980-2016. The most recent daily data (released in mid 2017) used in this study is version 3, retrieved from https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1328 (Thornton et al. 1997; Thornton et al. 2017). Daymet uses daily data from the Global Historical Climatology Network – Daily (GHCND) stations and the Community Collaborative Rain, Hail and Snow Network (CoCoRaHs) stations. Daymet fits a smooth surface in *x*, *y*, *z* using nearby stations. The grid cells are weighted using truncated Gaussian filter centered around the grid point, with the filter radius varying continuously in accordance with varying station density.   It is important to note that Daymet utilizes an extrapolation smoothing technique where the smoothing weight for no precipitation days is zero, resulting in values as weighted averages of daily precipitation events. This extrapolation smoothing technique is implemented to avoid a continuous drizzle bias seen with most smoothing techniques of daily precipitation in gridded datasets (Thorton et al. 1997).

For more information on the algorithms and revisions of the Daymet daily data please visit https://daac.ornl.gov/DAYMET/guides/Daymet_V3_CFMosaics.html.

2.6 Newman

The Newman data (Newman et al. 2015) is an observation-based ensemble of daily gridded datasets over the contiguous U.S., northern Mexico, and southern Canada with temperature and precipitation variables for the period 1980-2012. The probabilistic quantitative precipitation and temperature estimation is achieved through ensemble variance. Within the Newman dataset, there are 100 ensemble members, each with a slightly modified interpolation of temperature and precipitation. To generate each ensemble member, spatially correlated random noise was added with the purpose of increasing simulating realistic variability away from stations, where precipitation would otherwise be underestimated. Only a single ensemble member was used here, "conus_ens_001.nc". Although the 12-km resolution of the Newman data is the coarsest in this study, it was included in the study because it is independent from other datasets in that it doesn't scale to PRISM climatology like some of the others. Its unique methods of gridded, which were designed for the specific purpose of improving the representation or precipitation variability, also make it important to include within our study.

2.7 Other Datasets

Other datasets were analyzed, but ultimately not included in the results of this study. The first is the PERSIANN-CDR dataset (PERSIANN; Soroosh et al. 2014, Ashouri et al. 2015). PERSIANN provides daily precipitation amounts, estimated by GridSat-B1 infrared satellite data. Thus, PERSIANN is the only satellite-based gridded dataset within the study, making it important to compare the distinctions across the remaining datasets. However, the differences between PERSIANN and the other gridded datasets were ultimately too great to include within the final analysis. There were too many temporal, spatial, and resolution limitations, which lead to the decision to omit this dataset. Firstly, available data for PERSIANN only spanned 1983-2010, inconsistent with the normal 1981-2010 period analyzed within all other datasets. The spatial coverage was inconsistent, with many daily precipitation days missing due mechanical difficulties of the satellite's daily measurements. Finally, the dataset's resolution of 0.25° (approximately 27-km) was rather coarse compared to all other datasets. This was especially obvious within the daily precipitation patterns, as the Sierra Nevada mountains were not distinguishable, unlike all other datasets (Fig 2; pg 20). With these limitations and differences PERSIANN was ultimately discarded from final analysis. The second dataset was a secondary ensemble of the available Newman data. Due to the availability of multiple ensembles from the Newman data, it is important to note the variation between the different ensemble members. Therefore, "conus_ens_002.nc" was also analyzed to compare any distinctions (denoted as Newman (002) in Figure 2). The error metrics, precipitation statistics and extremes, and daily progression of precipitation amounts were examined. For this analysis, the results of these comparisons show that the secondary ensemble data –Newman (002)– were not too different to treat the separate ensemble members of Newman as completely independent datasets. Therefore, only the original ensemble member, denoted as "Newman" pertaining to the "conus_ens_001.nc" ensemble, is displayed in the Results section of this paper. Figure 2 best depicts the differences/inconsistencies of PERSIANN and the similarities of Newman (002), leading to decision to exclude both datasets from final analysis. A more detailed analysis on the storm depicted in Figure 2 is explained the Results section of this paper.

Dataset Name	WRF	Livneh	PRISM (AN81d)	Daymet	Newman (conus_ens_001.nc)
Category/Type	Dynamically downscaled reanalysis	Station-based	Station-based	Station-based	Station-based
Citation	Walton et al. 2018	Livneh et al. 2015	Daly et al. 2008	Thornton et al. 2017	Newman et at. 2015
Data available from	http://research.atmos.ucla.edu/csrl/data/	ftp://ftp.hydro.washington.edu/pub/blivneh/	http://prism.oregonstate.edu	https://daymet.ornl.gov	https://www.earthsystemgrid.org/dataset/gridded_precip_and_temp.html
Native Resolution	9-km	1/16° (~6 km)	~4-km (2.5 min)	1-km	12-km
Time Period	1980-2015	1915-2013	1981-2017	1980-2016	1980-2012
Input Data for Precipitation	NARR	COOP	COOP, WBAN, SNOTEL, RAWS, CDEC, Agrimet, & others	GHCN-Daily (Menne et al., 2012), and CoCoRaHs (https://www.cocorahs.org)	GHCN-Daily, and SNOTEL (from Natural Resources Conservation Service (NRCS) Snowpack Telemetry)
Downscaling/ Interpolation Method	WRF coupled to Noah-MP	SYMAP algorithm: inverse distance weighting with directional adjustment	Elevation-regression model with stations weighted based on multiple physical factors	Truncated Gaussian filter combined with elevation-regression	Probabilistic quantitative precipitation and temperature estimation is achieved through ensemble variance

Table 1: Individual Dataset Details. Specific attributes of each gridded dataset utilized within study.

3. Methods

3.1 Example Storm

A great deal of information regarding timing, magnitude and spatial smoothness of each dataset's precipitation field can be understood by looking at the daily variability during a particular storm. Therefore, a storm that passed through the domain, distributing a large amount of precipitation for many regions over a few days, was analyzed for comparison. Dozens of storms which passed through the area were assessed, and the storm highlighted in this paper – a storm from February 5-10, 1985, shows the daily distribution of precipitation amongst all the datasets –Newman (002) and PERSIANN included (Fig 2). This storm is a typical example of the kind of high precipitation events in California. This 6-day window was chosen intentionally, so as to investigate how each dataset differs (spatially, temporally, and quantitatively) in its development of daily precipitation to areas within our region.

3.2 Precipitation Statistics

To compare the five gridded datasets to station data, first extreme statistics were calculated for each gridded dataset, and then nearest grid cell was selected to compare with the station data. Thirteen indices of extreme precipitation were computed at each grid cell for every dataset. Most of these indices come from CCl/CLIVAR/JCOMM Expert Team (ET) on Climate Change Detection and Indices (ETCCDI) (https://www.clivar.org/organization/etccdi/etccdi.php.; http://www.climdex.org/indices.html).

Statistic	Units	Description/Comments
Rx1day	mm	Maximum 1-day precipitation value for time-period analyzed (either annual or record)
Rx5day	mm	Maximum consecutive 5-day precipitation value for the time-period analyzed (either annual or record)
SDII (simple day intensity index)	mm/day	For precipitation, it is defined as the ratio of total rainfall during the period analyzed to the number of days when rainfall occurred –generally regarding daily amounts ≥ 1 mm as wet days (https://gmao.gsfc.nasa.gov/research/subseasonal/atlas/Pindices-html/SDII.html)
R10mm	days	Total count of days when precipitation amount exceeds 10 mm during the time-period analyzed
R20mm	days	Total count of days when precipitation amount exceeds 20 mm during the time-period analyzed
R1mm	days	Total count of days when precipitation amount exceeds 1 mm during the time-period analyzed
Wet Day Frequency	days	Total count of wet days; i.e., when precipitation values are ≥ 1 mm
Dry Day Frequency	days	Total count of dry days; i.e., when precipitation values are < 1 mm
CDD	days	(Consecutive Dry Days is the maximum length of a dry spell; i.e., maximum number of consecutive dry days within the time-period analyzed. A dry day is defined as any daily precipitation amount < 1 mm
CWD	days	(Consecutive Wet Days is the maximum length of a wet spell; i.e., maximum number of consecutive wet days within the time-period analyzed. A wet day is defined as any daily precipitation amount ≥ 1 mm
Annual Mean Precipitation (AMP)	mm/day	The average precipitation amounts over all the days in the time-period analyzed (1981-2010)
R95pTOT	mm (or %; see text)	Defined as the total accumulated precipitation that fell above the 95th percentile
R99pTOT	mm (or %; see text)	Defined as the total accumulated precipitation that fell above the 99th percentile

Table 2: Precipitation Statistics Descriptions. Specific attributes of each precipitation statistic utilized within this study.

Table 2 gives the indices calculated in this study, their abbreviations, and descriptions. The RXXpTOT variables were calculated only on wet days over the 1981-2010 period. Each of these percentile statistics is a simple accumulation amount (usually in thousands of mm for the 30-year period). To gain better insights from the statistics, these will be presented as percentages, denoting what percentage of precipitation fell above the XX percentile. Each statistic is calculated in one of two ways: either as a "record" value indicating that it is calculated once over the entire 1981-2010 time period, or as an "average" in which case the statistic is calculated every year and then average value of these years is reported. For example, "Rx1day Record" yields the single largest daily precipitation amount within the entire 1981-2010 period. However, "Rx1day Average" denotes the average of the largest daily precipitation amount taken over each year in 1981-2010. In some cases, like wet day frequency or annual mean precipitation, the results are equivalent, so the distinction is unnecessary. The above extreme indices are calculated at each grid cell and are displayed in this paper in two ways: as differences and as percent differences relative to the GHCND station data. As station data are the best primary source data for precipitation, if a gridded dataset is consistently different from station data, then it is considered a strong indication of a bias.

3.3 Error Metrics

To distill these statistics into a succinct and digestible format, as well as to compare across datasets, two error metrics were used: mean absolute error (MAE) and bias, each expressed as a percent of the average value of that statistic. The mathematical formulas for these metrics are,

$$Mean\ Absolute\ Error = \frac{\sum abs(S_{Gridded\ Dataset} - S_{GHCND})}{\sum S_{GHCND}} \times 100$$

$$Bias = \frac{\sum (S_{Gridded\ Dataset} - S_{GHCND})}{\sum S_{GHCND}} \times\ 100,$$

where S represents an extreme precipitation statistic, and the summations are taken over all 221 stations in California. The MAE reveals how close, on average, a gridded dataset is to the station data and the bias tells us whether a gridded dataset tends to overestimate or underestimate that statistic in California, giving a sense of directionality of the errors, not present in the mean absolute error metric alone.

4. Results

4.1 Example Storm

Figure 2 shows the daily precipitation amounts from a storm that went through California February 5-10, 1985. Each dataset provides a different spatially complete picture of precipitation. First, the station data seem to have a sharper gradient for the allotment of daily precipitation compared with the gridded datasets. As the datasets' intrinsic goal is to make a spatially complete picture from spatially incomplete data, this may be somewhat expected, however, the differences in daily precipitation magnitudes are rather large. The timing of precipitation varies between gridded datasets. In particular, Livneh and Newman (especially ensemble 002), show as much as 30 mm of precipitation on February 6th in the parts of Northern California, while others are almost completely dry (Fig 2). These timing differences are also evident on February 8th, when Livneh and Newman (both ensemble members) have precipitation in Southern California, while the other datasets don't have precipitation there until the next day. On the 9th of February, when the storm appears to be moving out of the area, PRISM and Newman (particularly ensemble 002) still show a considerably larger amount of precipitation over the central Sierra. The proceeding date (February 10th) also shows extreme variation. All other datasets (including the station data) show no precipitation through California, while PRISM allots approximately 20 mm of precipitation for most of coastal Southern California. Indeed, these timing differences are not just limited to this storm are prevalent throughout the time period. It is theorized that this "pre-allotting" of precipitation a day early is due to the timing of the observations from Livneh's parent station data; which are somewhat irregular.

Daymet also appears to have more sharp cut offs of precipitation compared to the other gridded datasets. On this same day –i.e., when the bulk of the storm's precipitation occurs, the magnitudes are also rather distinct. Daymet's magnitudes vary smoothly where precipitation occurs, but abruptly drops to zero elsewhere. This pattern exists as a result of Daymet's interpolation method which includes the fitting of a smooth curve in (x,y,z) through the data, and explicitly computing the probability of precipitation occurrence, applying a threshold of 50% to determine whether precipitation occurs.

For this storm, WRF and Newman appear to have the peak daily precipitation values, yet WRF's highest magnitudes appear in the Sierras and Newman's (both ensembles) highest magnitudes appear near the coast. The magnitudes of Livneh, Daymet, and PERSIANN are rather muted compared to the others. However, due to the fact that Livneh's precipitation regime began a day earlier, the overall total from the storm isn't too dissimilar to the others' totals (Did you calculate totals? If so, you should plot it, maybe at the end? Otherwise, I wouldn't talk about totals). Effects of the Sierra Nevada and other topographic features are not evident in PERSIANN. This may be because it is a satellite-based product, or because it's resolution of approximately 27-km is rather coarse. Either way, it is obvious that, unlike the other datasets, it does not capture the orographic enhancements of precipitation that are a defining characteristic of precipitation in California.

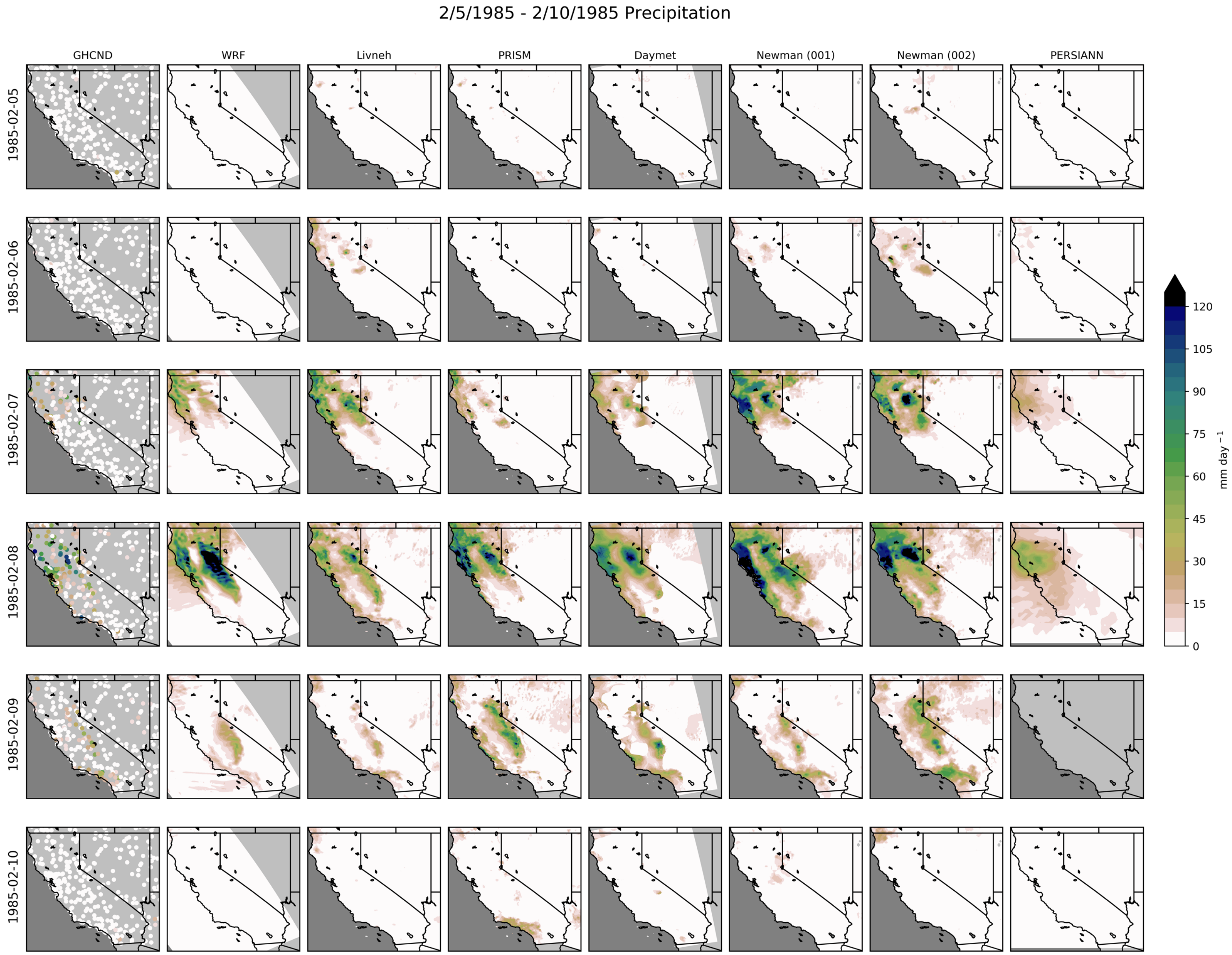

Figure 2: Example Storm Daily Progression. This figure shows the differences between the datasets for a storm that moved through the area of interest in February 1985. Each row shows the daily precipitation values for each dataset; GHCND, WRF, Livneh, PRISM, Daymet, Newman, Newman: secondary ensemble, and PERSIANN (satellite dataset), starting on February 5, 1985 and ending on January 10, 1985.

4.2 Precipitation Statistics

Thirteen different extreme precipitation statistics were calculated. With so many statistics, we focus on a representative subset: annual mean precipitation (AMP), precipitation amount on the wettest day (Rx1day Record), fraction of total precipitation from the wettest days (R95pTOT; as a percent), wet days per year, and consecutive dry days (CDD). For each statistic (plotted in figures 3-7), the top row of each plot displays the value of the statistic in each dataset. The second row shows the differences with GHCND at the station locations. The second row also shows the value of the statistic's two error metrics (MAE and bias) in the top right corner of each plot. The third row shows the percent differences of the GHNCD values at the station locations, while also including the average percent difference in the top right corner of each plot.

4.2.1 Annual Mean Precipitation (AMP)

AMP varies significantly with topography (Fig 3 Row 1). Most of the station-based datasets are within ~0.25 mm/day of the station values (Fig 3, row 2). Spatially, the pattern amongst all datasets, is very similar, with the largest disagreements in magnitude in high precipitation areas. WRF appears to underestimate slightly in the driest regions (such as the Mojave Desert) and overestimate slightly in the wettest (the high peaks of the Northern Sierra Nevada). Meanwhile, Newman consistently overestimates this statistic (bias of +17%), with percent differences >50% across the drier areas north and east of the moutnains in Southern California and east of the Sierra Nevada. Newman and WRF, both have similar MAEs of 0.4 mm/day (approximately 25%). The consistent overestimation in Newman is likely due to the spatially correlated random noise added to the precipitation field. WRF's small bias of -6% reflects offsetting errors, as WRF underestimates coastal precipitation and overestimates precipitation in the Northern Sierra Nevada. In contrast, Livneh and PRISM perform better match station data, with MAE of 12% (0.2 mm/day) and 11% (0.2 mm/day), respectively, and biases of 5% and 6%, respectively. Qualitatively, Daymet's results are consistently wetter than station data, especially in semi-arid regions east of the Sierra Nevada, and north and east of mountain ranges in Southern California. This could be to Daymet's desire to fit a smooth curve through the data, which may not catch the abrupt drop changes in precipitation going from the windward to the lee side of these mountain ranges. Overall, it appears that the four station-based gridded datasets slightly to moderately overestimate AMP, while WRF underestimates AMP. That the station-based datasets differ from the stations is somewhat surprising, since they are trained on the station data.

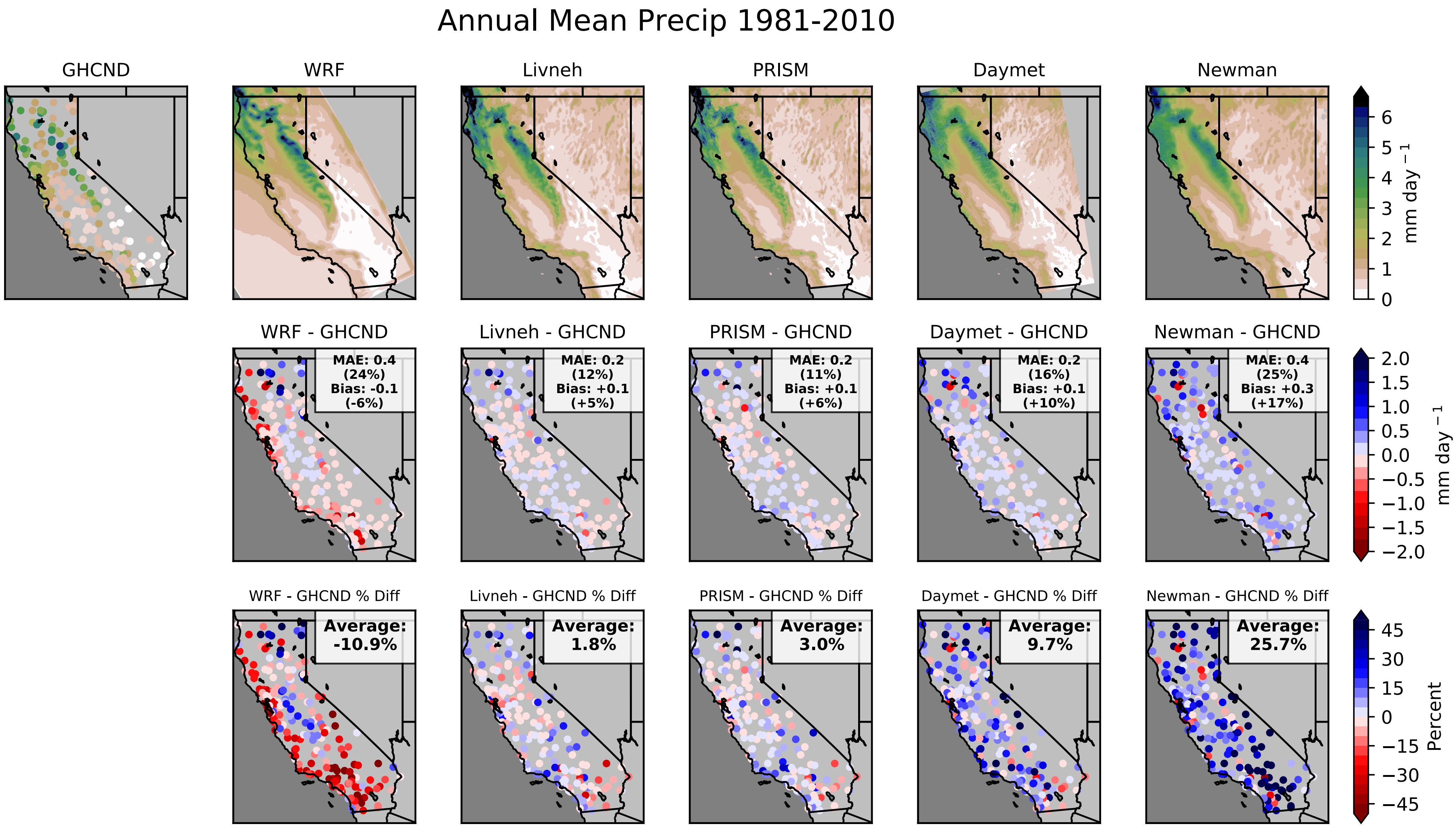

Figure 3: Annual Mean Precipitation (AMP). Top Row: Annual Mean Precipitation (1981-2010) for all 5 datasets. Middle Row: Quantitative difference between datasets regridded onto the GHCND data. The Mean Absolute Error (MAE) and Bias raw values and percent values are also shown in the top right of each plot; raw value units are the same as middle row units. Bottom Row: Same as Middle Row, but in percent difference. The average percent difference is also shown in the top right of each plot.

4.2.2 Rx1day (Record)

Rx1day Record (henceforth Rx1day) is the maximum single-day precipitation amount over the entire 1981-2010 period (Fig 4). Rx1day shows very significant differences among the datasets. WRF simulates more precipitation in on the windward side of the Sierra Nevada mountains. This could be due to the methodology applied to WRF's precipitation. Being the only dynamically downscaled dataset, the forcing applied to the boundaries create a spatially complete picture, precipitation magnitudes may be more sensitive to the highest topography of region. There is good reason to believe that WRF precipitation could be higher on the windward side, since the orography forces ascent of air masses for these storms. PRISM shows peculiar high magnitude values of highest single day precipitation in Northeastern California, near the King mountain range. It is speculated these high magnitudes are seen in this area because PRISM includes many different networks, many with stations in the King mountain range, leading to the notoriously high precipitation values of this area. Furthermore, unlike Daymet, which includes a similarly large number of stations, PRISM doesn't employ a surface smoothing technique to create a smooth precipitation surface. Thus, PRISM is able to capture this important peak. Daymet has the smoothest and smallest maximum single-day precipitation values, especially at high elevations. Overall, WRF, Livneh, PRISM, and Daymet all underestimate Rx1day, exhibiting dry biases: -12% for WRF, -17% for Livneh, -9% for PRISM, and -16% for Daymet; and underestimating station data by 20-50% at many stations. In stark contrast, Newman shows high values (>120 mm) over a much wider area compared to other datasets, especially near the Mojave Desert and east of the Sierra Nevada. It is not surprising that most station-based gridded datasets underestimate Rx1day. Typically, interpolation algorithms use a weighted average or regression function using multiple station locations. Thus, the value at any given grid cell is a weighted average over multiple station locations. So, if a station is experiencing its all-time maximum value, it will most likely be averaged out with slightly lower values from surrounding stations, producing a gridded value that is lower than the closest station. In stark contrast to the others, Newman frequently overestimates Rx1day, with a wet bias of +25% (CA average: +28.3 mm). It's not that Newman's peak values are higher but that high values extend over a much wider area than the other datasets, especially in areas like the Central Valley, the Mojave Desert, east of the Sierras. Newman exceeds station values by 50% at many stations in these regions, a strong indication that Newman is biased here. It seems highly likely that the overestimation of this extreme statistic is a result of adding spatially correlated random noise, which allows a grid cell value to exceed its nearest station.

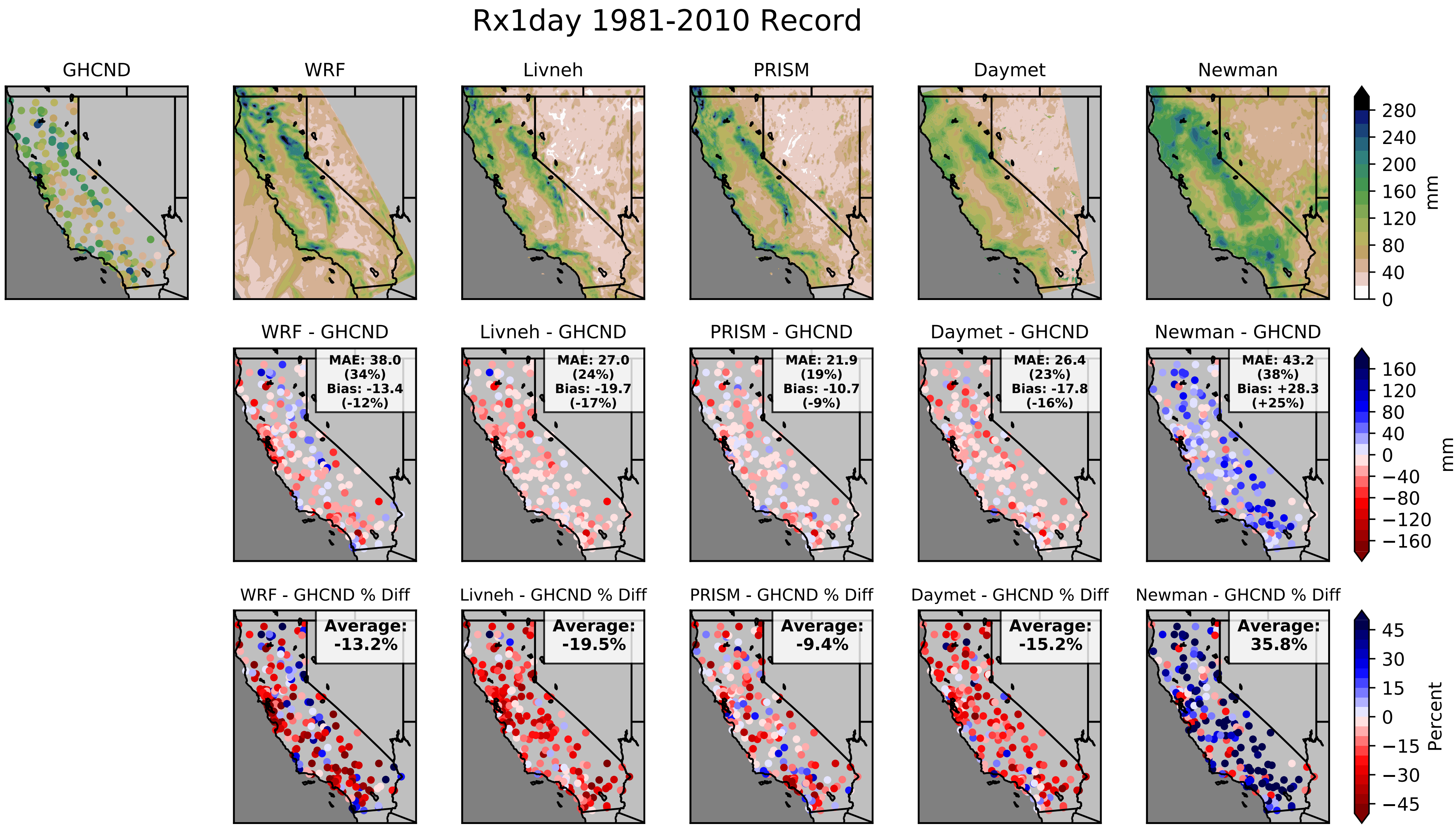

Figure 4: Rx1day Record. Top Row: Maximum daily precipitation amount for the entire time period (1981-2010) for all 5 datasets. Middle Row: Quantitative difference between datasets regridded onto the GHCND data. The Mean Absolute Error (MAE) and Bias raw values and percent values are also shown in the top right of each plot; raw value units are the same as middle row units. Bottom Row: Same as Middle Row, but in percent difference. The average percent difference is also shown in the top right of each plot.

4.2.3 R95pTOT (Percentages)

R95pTOT, the amount of precipitation accumulated on days above the 95th percentile, expressed as a percentage of total rainfall, varies considerably across the five datasets (Fig 5). Daymet consistently underestimates R95pTOT (MAE: 12%, bias: -10%), while Newman overestimates R95pTOT (MAE: 22%, bias: +22%). In Southern California, Newman shows that more than 35% of all precipitation fell from extreme days, a much higher fraction than station data suggest (typically around 20-25%, see Fig 5 Row 1). The spatial pattern of Newman in Southern California is also different than the others, with high values east and north of the Southern California mountain complexes that are not supported by the station data. In fact, Newman overestimates R95pTOT by a factor of 1.5 at stations in these areas. Unlike Daymet and Newman, which uniformly under and overestimate R95pTOT, respectively, the other datasets, WRF, Livneh, and PRISM, have smaller differences with station data, underestimating in some areas and overestimating in others. Errors are especially small for Livneh (MAE: 6%, bias: <1%) and PRISM (MAE: 6%, bias: +2%). For WRF (MAE: 11%), errors in R95pTOT are topographically dependent, with the windward sides of the Sierra Nevada and coastal mountain ranges seeing the largest overestimations (by a factor of approximately 2).

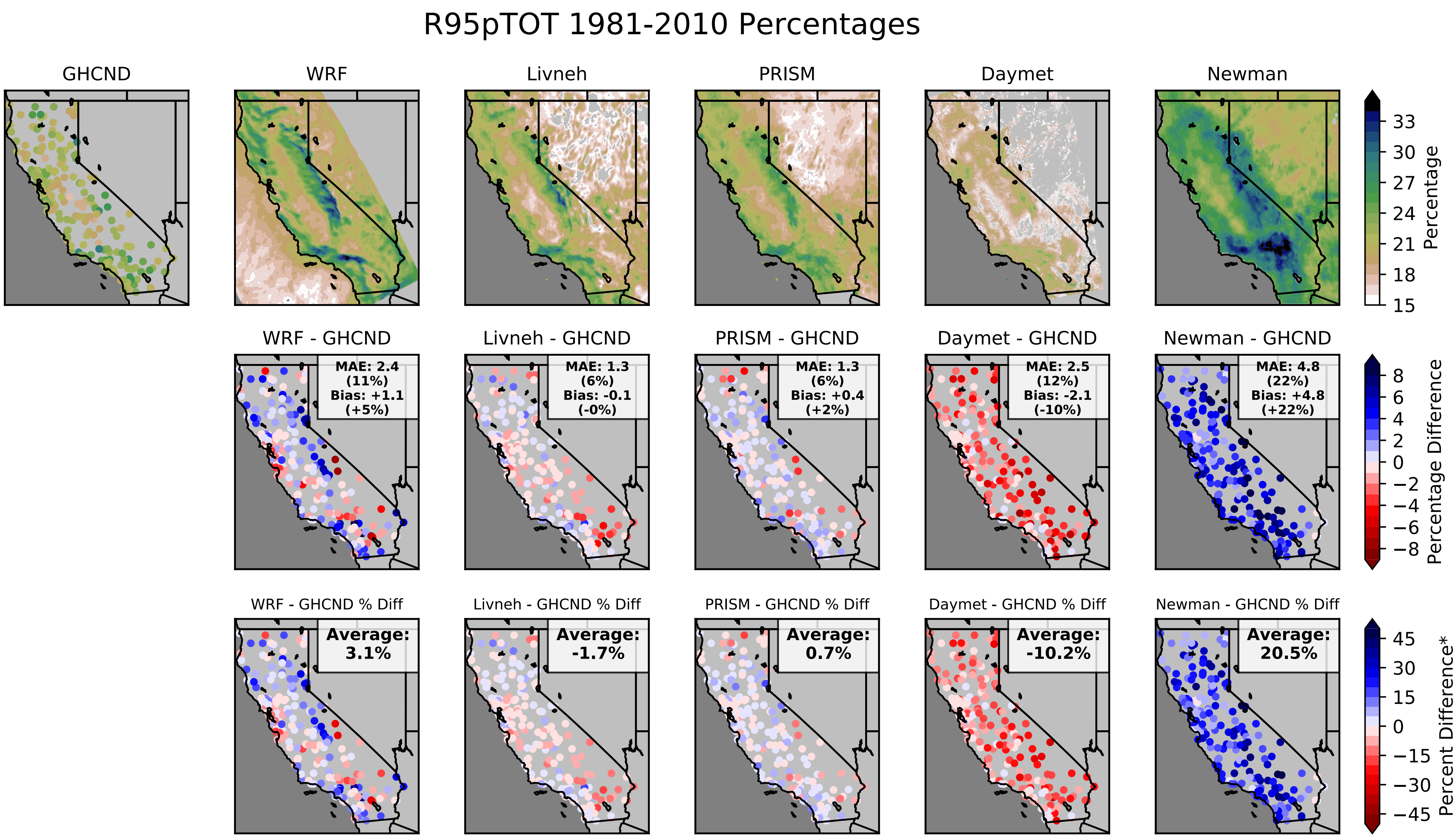

Figure 5: R95pTOT (Percentages). Top Row: Percentage of precipitation that fell above the 95th percentile of total precipitation for all 5 datasets. Middle Row: Quantitative difference between datasets regridded onto the GHCND data. The Mean Absolute Error (MAE) and Bias raw values and percent values are also shown in the top right of each plot; raw value units are the same as middle row units. Bottom Row: Same as Middle Row, but in percent difference. The average percent difference is also shown in the top right of each plot.

*Note: Middle row refers to the quantitative difference in percentages, while the bottom row calculates regular percent differences for variable R95pTOT.

4.2.4 Wet Day Frequency

All datasets overestimate the number of wet days (daily precipitation ≥ 1 mm; Fig 6). Still, there are significant differences between the datasets in the locations and magnitudes of these overestimations. WRF agrees most closely with the station data (MAE: 14%, bias: +10%). Daymet agrees the best of the station-based datasets (MAE: 18%, bias: +17%) and, like WRF, does a better job of capturing the low frequency of precipitation in Southern California (< 30 wet days per year). Livneh, PRISM, and Newman all overestimate wet days in Southern California. Livneh has the largest domain-wide overestimation (MAE: 35%, bias: +35%), with Newman a close second (MAE: 34%, bias: +34%), and PRISM third most (MAE: 18%, bias: +17%). These datasets have especially large overestimations in the coastal mountains of Northern California and the Sierra Nevada, where their values can exceed the stations' by more than 30 wet days per year. The overestimation results across all gridded datasets are somewhat expected, as many gridded datasets are known to have a "drizzle problem", whereby the datasets record trace or small amounts of precipitation (drizzle) more frequently than observations suggest (Ensor and Robeson 2008). Within gridded datasets, this known "drizzle" issue is hypothesized to arise from the averaging between various stations some with and some without precipitation, resulting in small values assigned to many grid cells throughout the domain.

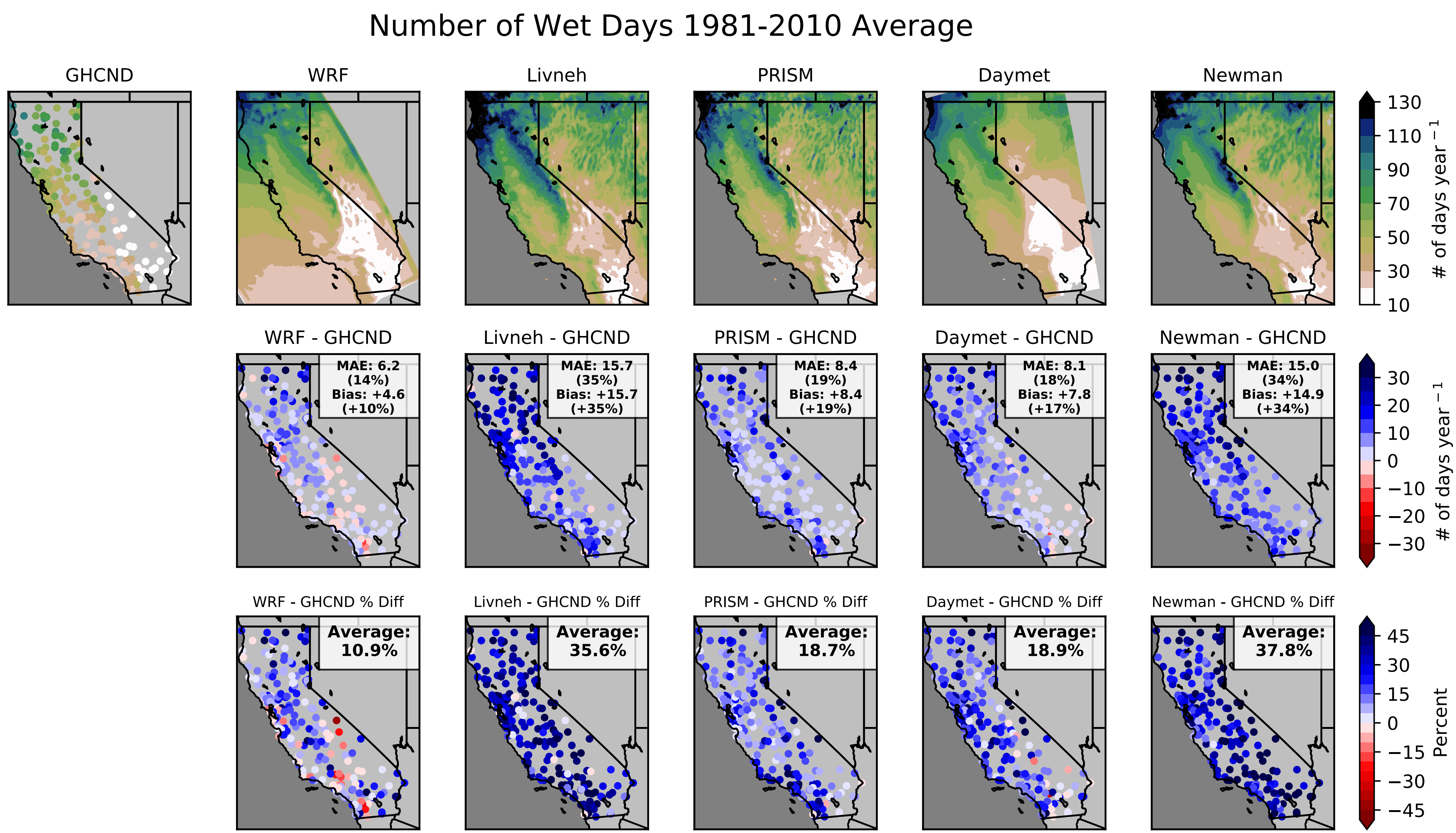

Figure 6: Number of Wet Days. Top Row: Annual number of Wet Days (≥1 mm) (averaged over 1981-2010) for all 5 datasets. Middle Row: Quantitative difference between datasets regridded onto the GHCND data. The Mean Absolute Error (MAE) and Bias raw values and percent values are also shown in the top right of each plot; raw value units are the same as middle row units. Bottom Row: Same as Middle Row, but in percent difference. The average percent difference is also shown in the top right of each plot.

4.2.5 CDD (Consecutive Dry Days)

There are large variations in CDD (representing the average length of the longest continuous streak of daily precipitation amounts < 1 mm within a year; Figure 7) in California, with areas of Southern California averaging over 150 consecutive dry days a year, while Northern California and the Sierra Nevada average less than 50 per year. Among all the datasets, CDD agrees reasonably well with the station data, with no dataset having CA-average larger than 13 days/year (11%). In general, Livneh slightly, but consistently overestimates the longest dry streak (MAE: 8%, bias: +5%). WRF (MAE: 11%, bias: +3%) and Daymet (MAE: 9%, bias: +2%) also slightly overestimate in some places (like the Mojave Desert). Although WRF's bias is only +3%, this is the result of offsetting errors: WRF overestimates in some locations by up to 40% (like the Mojave Desert) and underestimates in others by up to 40%. In general, the biases and results of PRISM (MAE: 7%, bias: -1%) and Newman (MAE: 9%, bias: -2%) suggest very little systematic over or underestimation for longest dry streak. In total, these two datasets perform best in simulating consecutive dry days.

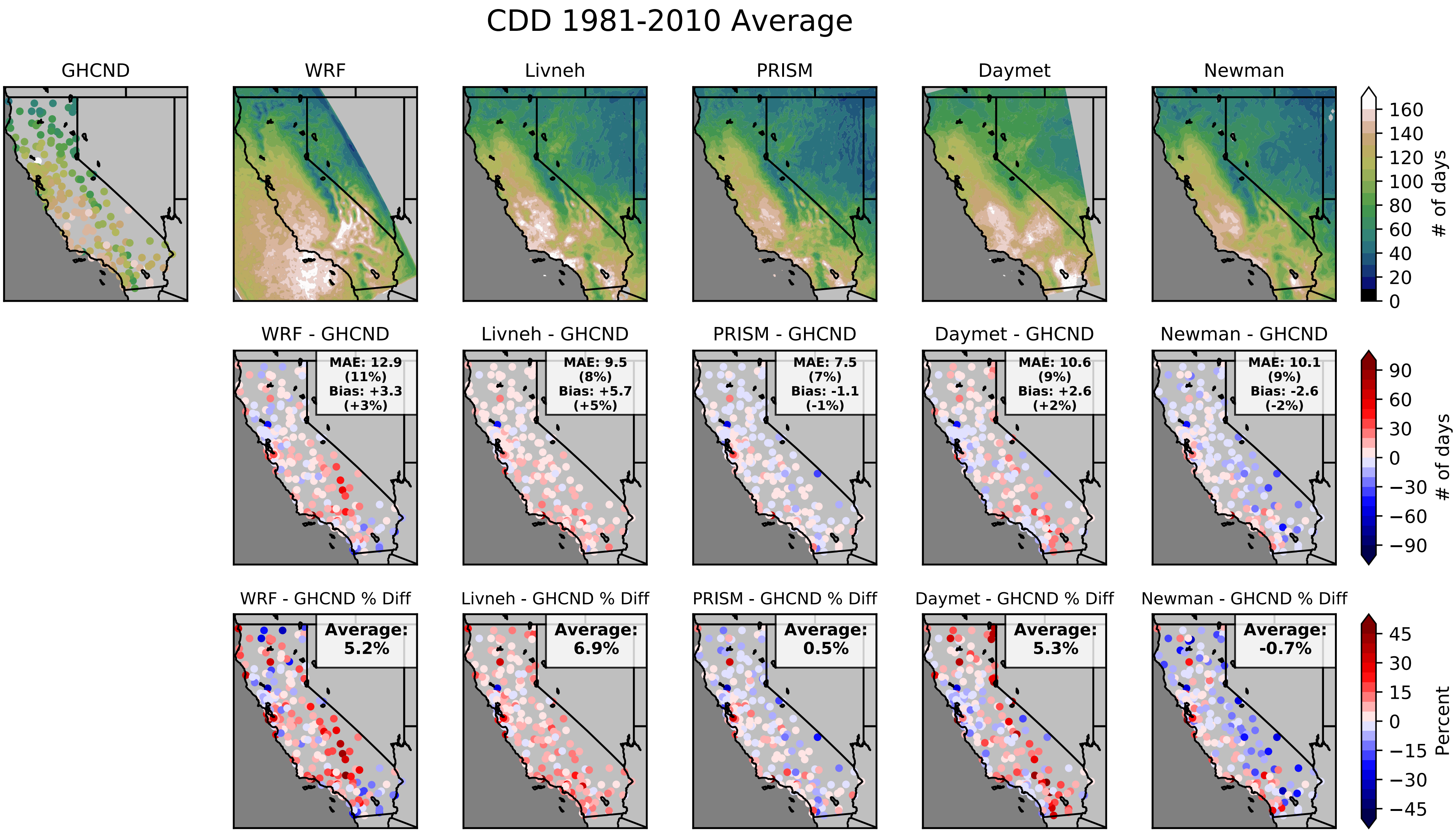

Figure 7: CDD (Consecutive Dry Days). Top Row: The average length of the longest continuous streak of daily precipitation < 1mm (averaged over 1981-2010) for all 5 datasets. Middle Row: Quantitative difference between datasets regridded onto the GHCND data. The Mean Absolute Error (MAE) and Bias raw values and percent values are also shown in the top right of each plot; raw value units are the same as middle row units. Bottom Row: Same as Middle Row, but in percent difference. The average percent difference is also shown in the top right of each plot.

*Note: the middle and bottom row's shading has been flipped to continue the notion of blue as too wet and red as too dry.

4.3 All Statistics/Extremes

Figure 8 shows the MAE and bias for all statistics as a percentage of the California-average value of that statistic. Across all statistics, MAEs range from 6%-54% and biases range from -19% to 54%. Differences in Annual Mean Precipitation between the gridded datasets and the station data (MAEs ranging from 11 to 25%, biases ranging from -6 to +17%) were larger than expected. Given that the station-based datasets are trained on the station data and the fact that AMP is average quantity, one would expect this variable to closely match the station data, yet MAE exceeded 10% for all station-based datasets. The remaining dataset, WRF, has a relatively low AMP bias of -6%. Being on par with station-based datasets, this is a rather impressive feat for the dynamically-downscaled dataset.

The Rx1day Record and Average MAE values differ from the station data by an average of approximately 26% and 22%, respectively, while their biases differ by an average of by an average of about +14% –across all datasets, save Newman– showing a clear underestimation in maximum daily precipitation. Newman overestimates this statistic by +25% and +19%, respectively. These four overestimating results were also expected as per the reasoning discussed previously; the datasets generally have set maximum precipitation amounts, in which the algorithms determining precipitation do not allow daily accumulations to exceed. This suggests that the maximum single day precipitation is often a very difficult extreme statistic to produce within gridded datasets.

The extreme precipitation distribution statistics, R99pTOT and R95pTOT, are well simulated for most of the datasets, except Newman and Daymet. While most datasets consistently overestimate the amount of extreme precipitation, Daymet appears to underestimate these statistics by -13% and -10%, respectively. This suggests that much of the precipitation from Daymet occurs on non-extreme precipitation days. This conclusion is also substantiated by the Rx1day and Rx5day results, with Daymet having the second greatest underestimations across the datasets. Newman (again) appears to be the outlier, overestimating these statistics by +31% and +22%, respectively.

Across all datasets, no matter the interpolation method, frequency of wet days is overestimated by as little as +10% (bias from WRF) to as much as +35% (bias of Livneh). With the overestimated wet day frequency results, it is not surprising that CWD (consecutive wet days) is also overestimated, sometimes by severe amounts; +43% bias of Newman and +54% bias of Livneh. This may be exposing the known issue from gridded datasets, drizzle problem: precipitation allotment is often too freely given to grid cells due to averaging problems.

Overall, Newman appears to have the largest errors and consistently overestimates the amount of precipitation. When averaged over all of the statistics, Newman MAE of 30% and an average bias of +21%. The consistent overestimation of precip intensity from Newman may be due to the added spatially correlated random noise to the daily precipitation distribution. On the other hand, PRISM appears has the lowest average MAE of 14% and an average bias of +4%. In all, it is important to remember that although users may assume that gridded datasets are as trustworthy as the station data they are generated from, our results, however suggest that gridded datasets differ significantly from station data in many respects.

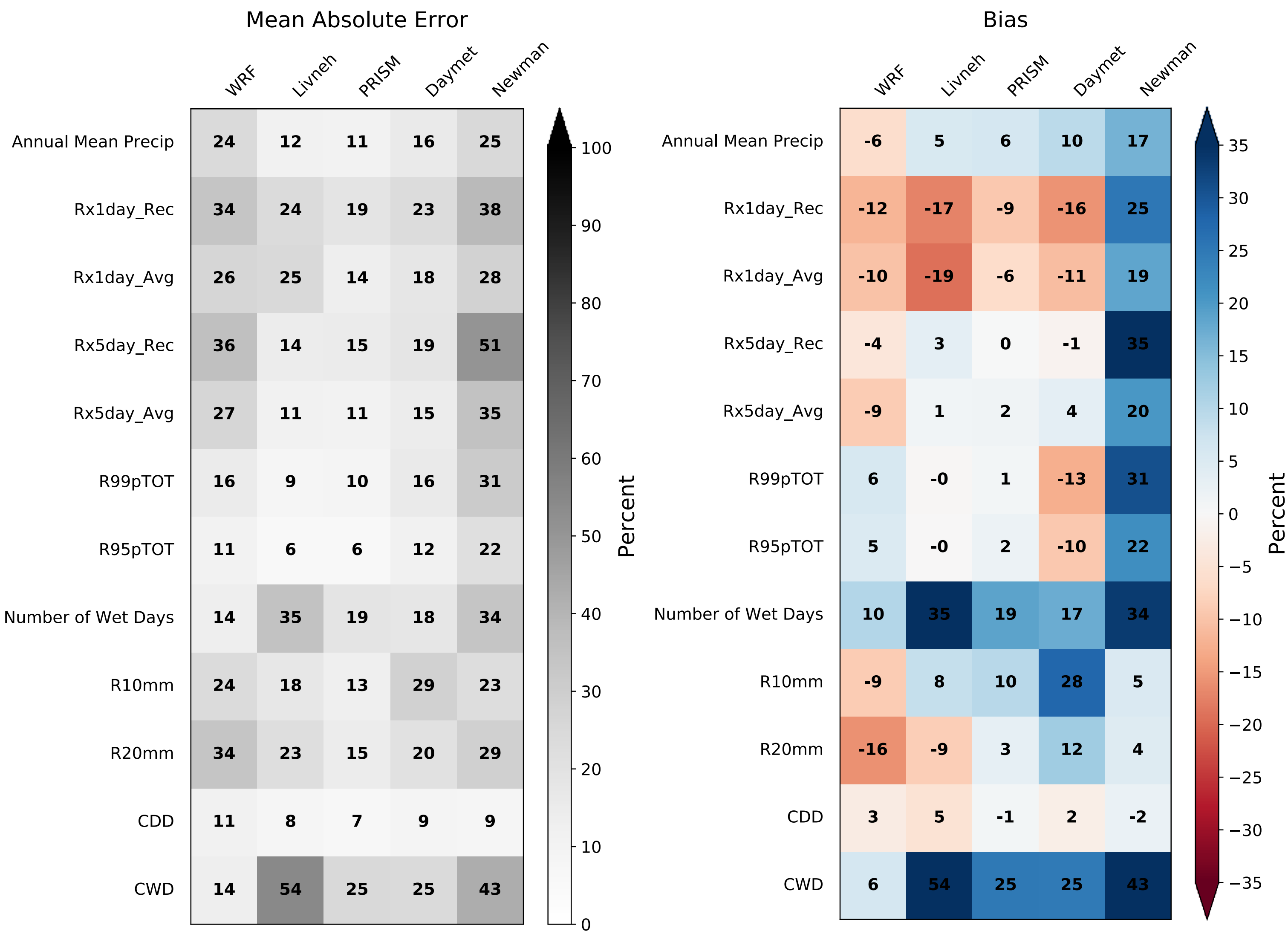

Figure 8: Mean Absolute Error (MAE) and Bias Metrics. (LEFT) The mean absolute error for all gridded datasets compared to GHCND California stations. The shading and number reflect the percentage difference for each particular statistic; 0 being perfect. (RIGHT) The bias for all the gridded datasets compared to GHCND California stations. The shading and number reflect the percentage difference for each particular statistic; red being an underestimation (too dry) and blue being an overestimation (too wet).

*Note: the shading for CDD has been flipped to continue the notion of blue as too wet and red as too dry.

5. Discussion and Conclusions

5.1 General Gridded Dataset Biases and/or Errors

In total, this assessment of extreme precipitation statistics shows gridded datasets differ – sometimes greatly– from each other and from station data. The regridded datasets onto station data grids give quantitative differences –which can be thought of as the extreme precipitation errors and/or biases– and show clear overestimations or underestimations that can be tied directly back to interpolation methodology. In identifying such errors and the reasoning behind them, hopefully these results can be useful in selecting the appropriate dataset for the task at hand, for identifying areas where the datasets could be improved, and for creating new ones. Overall, PRISM has the least bias and smallest errors, giving it the highest marks. However, it may not be the best for each individual statistic.

All of the gridded datasets (save Newman) tend to *underestimate* the intensity of precipitation extremes. This result was not surprising for the station-based datasets, due to the interpolation methods used. Typically, interpolation algorithms use a weighted average or regression function of multiple nearby stations to calculate the value at a grid point. Thus, the value at a grid point is essentially the weighted average of it nearest neighbors, meaning that it will nearly always be greater than the lowest nearby station value and less than the highest nearby station value. So, if a station is experiencing it's all time maximum—a value unlikely to be exceeded by another nearby station—the nearest grid point will have a less than that. This underestimation of precipitation intensity is consistent with the findings of Behnke et al. (2016a), which found large and systematic underestimation of precipitation at high intensities.

5.2 Individual Gridded Dataset Assessments

Newman is consistently too wet, overestimating precipitation amounts and frequency. This is consistent with findings of the dataset's creators, which noted generally good agreement except for precipitation extremes, specifically, events with precipitation magnitudes > 50 mm/day, for which a slight wet bias was identified (Newman et al. 2015). The addition of the spatially correlated random noise to daily precipitation amounts forces the distribution to become more normally distributed, intentionally altering an observed daily precipitation distribution that is highly skewed. Personal communications with the authors indicate that this is a known issue and that fixes are likely to be made.

Daymet appears to underestimate precipitation intensity. This is true for Rx1day, where it has muted precipitation in high elevation areas, but also in R99pTOT and R95pTOT, where it shows smaller fractions of precipitation coming from extremes. Since total precipitation is not underestimated, then Daymet must be compensating by having more precipitation on less intense days. produces more precipitation on its less intense days and less on the most extreme days than the other datasets. Indeed, the frequency of heavy and very heavy precipitation (R10mm and R20mm) are overestimated compared to the other gridded datasets. Therefore, it can be deduced that Daymet underestimates intensity on the most extreme days while overestimating the frequency of mid to high values. This could be the result of the smoothing nature Daymet's interpolation algorithm and its methods for determining precipitation occurrence. Haylock et al. (2008) and Hofstra et al. (2010) found that smoothing methods have a strong tendency to lead to daily precipitation and/or temperature values less than the "true" area average, also noting that the smoothing is most severe for higher percentiles, leading to underestimated extremes. Thus, their idea that underestimation of extreme precipitation due to smoothing is supported by this paper's findings. First, Daymet's creators believed in the importance of having spatially smooth temperature and precipitation fields; specifically stating, "We prefer an interpolation surface that is continuous, though we do not impose the condition that it be perfectly smooth, in that its first- and higher-order derivatives are allowed to be discontinuous" (Thornton et al. 1997). Second, for determining precipitation amounts, the interpolation method first determines the probability of precipitation occurrence for that location and date, and if the probability is less than an assigned threshold, the daily precipitation is assigned a value of zero. Also, the smoothed value for precipitation represents a weighted average of daily precipitation events, except on zero precipitation days; in which case the smoothing filter is also set to zero (Thornton et al. 1997). This was implemented to solve the "constant drizzle" problem which would arise from a standard smoothing filter applied to a precipitation time series. Daymet is the only dataset in this analysis to explicitly account for this known "drizzle" issue by calculating the probability of precipitation occurrence.

WRF is the only dynamically-downscaled dataset within this study. While the station-based gridded datasets start with the station data and try to create a spatially complete picture, WRF solves for the entire picture at once, using only forcings at the domain boundaries. For this reason, it is inherently different. One consistent way that WRF differs from the other datasets is that the direction of its errors varies in space: WRF may overestimate in one region while underestimating in another. Meanwhile the station-based datasets typically overestimate or underestimate over the whole state. For many statistics, WRF performs similarly compared to the station-based datasets. However, WRF has a dry bias in the coastal mountains. This may be due to WRF's relatively coarse 9-km resolution.

Livneh and PRISM are not only the most similar of the datasets, but also appear to be the best performing datasets when compared to the station data. Much of the similarity can be attributed to Livneh correcting its monthly climatology to that of PRISM. Unlike most of the other datasets, Livneh partitions daily precipitation totals from station between the current and previous day based on time of observation. Thus, single daily precipitation values may be spread over multiple days. This could explain why Livneh severely overestimates the number of wet days and consecutive wet days.

5.3 Overall Conclusions and Outlook

One reasonable question is whether using GHCND data to validate the gridded datasets is improper, since all of the station-based gridded datasets use many if not all of the GHCND stations as input. It would be difficult to achieve the extensive temporal and spatial coverage needed to evaluate extreme precipitation without GHCND. Indeed, one challenge of evaluating gridded datasets is that as more and more station networks are used in gridded datasets, it becomes harder to find independent datasets. For California, one possible option would be the CIMIS station (e.g. Behnke et al. 2016a), but it doesn't have nearly sufficient spatial coverage. In fact, one key result of this study is that most gridded datasets still underestimate precipitation extremes at station locations, despite being trained on that station data. This shows that interpolation methods applied to the gridded datasets are almost certainly the main causes for the corresponding errors and/or biases. It is possible that the nearest neighbor technique used to sample the gridded datasets at the station locations introduces errors as well, especially in areas with strong terrain gradients where a mismatch in elevation between the station and nearest grid cell could arise. One additional critique of evaluating against GHCND data is that interpolation methods that try to fit to the station data exactly are extended an advantage that they would not have when tested on an independent dataset. If this were the main cause of errors and/or biases, then –within this study– one would expect Livneh to give the absolute best results for the extreme statistics, due to the fact that it uses inverse distance weighting to generate its values; but it does not perform the best.

Although these results focused on California specifically, it is likely these errors and/or biases identified here would be present for other areas throughout the country and beyond. This due to the fact that the gridded datatsets' errors and/or biases can be at least partially attributable to interpolation methods, and not necessarily to the distinct California topography or climate. In particular, the underestimation of maximum single day precipitation is most likely to show similar results over the entire US and beyond, as the study shows this phenomenon to be a more systematic error, present in every dataset unless certain interpolation methods override this bias (i.e., Newman's added spatially correlated noise). In all, these results show that there are large differences between the gridded datasets as well as significant differences with station observations. This should serve as a strong reminder that gridded datasets are imperfect and should not be treated as absolute fact, but rather as the best spatially complete estimates of a complex and sometimes unknown picture.

6. References

Adam, J.C., Lettenmaier, D.P., 2003. Adjustment of global gridded precipitation for systematic bias. *Journal of Geophysical Research*. 108 (D9), 4257.

Ashouri, H., Hsu, K.L., Sorooshian, S., Braithwaite, D.K., Knapp, K.R., Cecil, L.D., Nelson, B.R. and Prat, O.P., 2015. PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bulletin of the American Meteorological Society*, *96*(1), pp.69-83.

Behnke, R., S. Vavrus, A. Allstadt, T. Albright, W. Thogmartin, and V. Radeloff, 2016a: Evaluation of downscaled, gridded climate data for the conterminous United States. *Ecological Applications*, **26**(5), 1338–1351. http://dx.doi.org/10.1002/15-1061.

Behnke, R., S. Vavrus, A. Allstadt, T. Albright, W. Thogmartin, and V. Radeloff, 2016b: Data from: Evaluation of downscaled, gridded climate data for the conterminous United States. *Dryad Digital Repository*, doi: http://dx.doi.org/10.5061/dryad.7tv80.

Caldwell, P. M., H.-N. S. Chin, D. C. Bader, and G. Bala, 2009: Evaluation of a WRF-based dynamical downscaling simulation over California. *Climatic Change*, 95, 499–521.

Cressman, G. P. 1959. An operational objective analysis system. *Monthly Weather Review* 87:367–374.

Daly, C., R. P. Neilson, and D. L. Phillips, 1994: A statistical–topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteor.*, **33**, 140–158, doi:10.1175/ 1520-0450(1994)033,0140:ASTMFM.2.0.CO;2.

Daly, C., 2006: Guidelines for assessing the suitability of spatial climate data sets. *Int. J. Climatol.*, 26, 707–721, doi:10.1002/joc.1322.

Daly C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, P. P. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol*., **28**, 2031–2064, doi: 10.1002/joc.1688.

Ensor, L. A., and S. M. Robeson, 2008: Statistical characteristics of daily precipitation: comparisons of gridded and point datasets. *J. Appl. Meteor. Climatol.,* 47, 2468–2476, doi:10.1175/ 2008JAMC1757.1.

Eum, H.-I., Y. Dibike, T. Prowse, and B. Bonsal, 2014: Inter-comparison of high-resolution gridded climate data sets and their implication on hydrological model simulation over the Athabasca watershed, Canada. *Hydrol. Processes*, 28, 4250–4271, doi:https://doi.org/10.1002/hyp.10236.

Førland E, Institutt NM. 1996. Manual for Operational Correction of Nordic Precipitation Data. Norwegian Meteorological Institute.

Gandin, L. S. 1965. Objective analysis of meteorological fields. Pages 242. Translated from Russian, Jerusa-lem, Israel Program for Scientific Translations, Gidrometeoizdat, Leningrad.

Gbambie A.S.B., Poulin A., Boucher M.A., Arsenault R. 2017. Added value of alternative information in interpolated precipitation datasets for hydrology. *J. Hydrometeorol.* 18: 247–264. https://doi.org/10.1175/JHM-D-16-0032.1.

Gervais, M., Gyakum, J. R., Atallah, E., Tremblay, L. B., & Neale, R. B. (2014). How well are the distribution and extreme values of daily precipitation over North America represented in the community climate system model? A comparison to reanalysis, satellite, and gridded station data. *Journal of Climate,* 27, 5219–5239. https://doi.org/10.1175/JCLI-D-13-00320.1

Goodison B.E., Louie P.Y., Yang D. 1997. The WMO solid precipitation measurement intercomparison. World Meteorological Organization-Publications-WMO TD, Report No. 67, 65–70.

Gutmann E.D., R.M. Rasmussen, C. Liu, K. Ikeda, D.J. Gochis, M.P. Clark, J. Dudhia, G. Thompson A comparison of statistical and dynamical downscaling of winter precipitation over complex terrain. *J. Clim.*, 25 (2012), pp. 262-281, 10.1175/2011JCLI4109.1.

Henn, B., Newman, A.J., Livneh, B., Daly, C. and Lundquist, J.D., 2018. An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of Hydrology,* 556, pp.1205-1219.

Haylock M, Hofstra N, Klein Tank A, Klok E, Jones P, New M. 2008. A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. *J. Geophys. Res. -Atmos.* 113(D20), doi: 10.1029/2008JD010201.

Hofstra N, New M, McSweeney C. 2010. The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data. *Clim. Dyn.* 35(5): 841–858, doi: 10.1007/s00382-009-0698-1.

Johns, C. J., D. Nychka, T. G. F. Kittel, and C. Daly. 2003. Infilling sparse records of spatial fields*. Journal of the American Statistical Association.* 98:796–806.

Journel, A., and C. Huijbregts. 1978. Mining geostatistics. Academic Press, New York, USA.

Legates, D.R., 1987. A climatology of global precipitation. Publications in climatology 40 (1), 86.

Livneh, B., E. A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K. M. Andreadis, E. P. Maurer, and D. P. Lettenmaier, 2013: A long- term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. *J. Climate*, **26**, 9384–9392, doi:10.1175/JCLI-D-12-00508.1.

Livneh, B., T. Bohn, D. Pierce, F. Muñoz-Arriola, B. Nijssen, R. Vose, D. Cayan, L. Brekke, (2015). A spatially comprehensive, meteorological data set for Mexico, the U.S., and southern Canada (NCEI Accession 0129374). Version 1.1. NOAA National Centers for Environmental Information. Dataset. doi:10.7289/V5X34VF6.

Mannshardt-Shamseldin, E.C., Smith, R.L., Sain, S.R., Mearns, L.O. and Cooley, D., 2010. Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data. *The Annals of Applied Statistics*, 4(1), pp.484-502.

Maurer, E. P., A. W. Wood, J. C. Adam, D. P. Lettenmaier, and B. Nijssen, 2002: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *J. Climate*, **15**, 3237–3251, doi:10.1175/ 1520-0442(2002)015,3237:ALTHBD.2.0.CO;2.

Menne, M. J., I. Durre, R. S. Vose, B. E. Gleason, T. G. Houston, 2012a: An overview of the Global Historical Climatology Network-Daily Database. *J. Atmos. Oceanic Technol.*, **29**: 897–910, doi:10.1175/JTECH-D-11-00103.1.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, doi:10.1175/BAMS-87-3-343.

Mizukami, N. and Smith, M.B., 2012. Analysis of inconsistencies in multi-year gridded quantitative precipitation estimate over complex terrain and its impact on hydrologic modeling. *Journal of Hydrology*, 428, pp.129-141.

Mote, P.W., A.F. Hamlet, M.P. Clark, and D.P. Lettenmaier. 2005. Declining mountain snowpack in western North America. *B. Am. Meteorol. Soc.* 86(1):39–49.

Newman A., M. Clark, J. Craig, B. Nijssen, A. Wood, E. Gutmann, N. Mizukami, L. Brekke, and J. Arnold, 2015: Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States. J. *Hydrometeor,* 16, 2481–2500. doi: http://dx.doi.org/10.1175/JHM-D-15-0026.1

Pierce, D. W., D.R. Cayan, B.L. Thrasher. Statistical downscaling using localized constructed analogs (LOCA) *J. Hydrometeorol.*, 15 (6) (2014), pp. 2558-2585.

Prein, A.F. and Gobiet, A., 2017. Impacts of uncertainties in European gridded precipitation observations on regional climate analysis. *International Journal of Climatology*, 37(1), pp.305-327.

Sevruk, B., 1982. Methods of correction for systematic error in point precipitation measurement for operational use Operational hydrology report, 21. Secretariat of the World Meteorological Organization, Geneva, Switzerland, 91 pp.

Shepard, D. S., 1984: Computer mapping: The SYMAP interpolation algorithm. *Spatial Statistics and Models*, G. L. Gaile and C. J. Willmott, Eds., D. Reidel, 133–145.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Technical Note, NCAR/TN-4751+STR.

Soroosh Sorooshian, Kuolin Hsu, Dan Braithwaite, Hamed Ashouri, and NOAA CDR Program (2014): NOAA Climate Data Record (CDR) of Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks (PERSIANN-CDR), Version 1 Revision 1. NOAA National Centers for Environmental Information. DOI:10.7289/V51V5BWQ.

Thornton, P. E., S. W. Running, and M. A. White, 1997: Generating surfaces of daily meteorological variables over large regions of complex terrain. *J. Hydrology*. **190**, 214–251, doi: 10.1016/S0022-1694(96)03128-9.

Thornton, P.E., M.M. Thornton, B.W. Mayer, Y. Wei, R. Devarakonda, R.S. Vose, and R.B. Cook. 2017. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 3. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1328.

Tozer, C., A. Kiem, and D. Verdon-Kidd, 2012: On the uncertainties associated with using gridded rainfall data as a proxy for observed. *Hydrol. Earth Syst. Sci.*, 16, 1481–1499, doi:https://doi.org/10.5194/hess-16-1481-2012.

Walton, D. B., F. Sun, A. Hall, and S. Capps, 2015: A hybrid dynamical–statistical downscaling technique. Part I: Development and validation of the technique. *J. Climate*, 28, 4597–4617, https://doi.org/10.1175/JCLI-D-14-00196.1.

Walton, D. B., and A. Hall 2018: An Assessment of High-Resolution Gridded Temperature Datasets over California. *J. Climate*, 31, 3789–3810, https://doi.org/10.1175/JCLI-D-17-0410.1.